DIMENSIONALITY REDUCTION

→ The problem of multicollinearity give rise to the biggest problem of data modeling - Overfitting.

- Reason of multicollinearity is having correlated features. Often, the amount of data that one gets can be overwhelming with a lot of features making the task of data exploration very tough.
- The high amount of data has its own disadvantages such as it increases the computational time, decrease the storage and most importantly causes multicollinearity causing the models to overfit.
- Dimensionality Reduction help us in reducing the processing time allowing us to perform more complex algorithms. Another benefit of having the data in low dimensions is that it frees storage space. The most important benefit of Feature Reduction is that it takes care of problem of multicollinear
- Thus Dimensionality Reduction helps in making the process of data analysis faster and more accurate.
- There are many techniques in Dimensionality Reduction —

1) Principal Component Analysis (PCA) —

- We transform our feature into a lower number of artifical features without losing much of the information.
- In this method, features are transformed into a set of 'artifical features'
- These 'artifical features' are known as Principal Components where the first component contains most of the information that can be contained in a single 'artifical feature' and we are left to select the number of components in order to reduce the features.
- Here the feature are not explicitly dropped rather the variation is extracted saving the loss of data.

2) Factor Analysis —

- In factor analysis, the feature are grouped based on their similarity which is determined by their shared variance, and then the user can pick relevant features from these groups making the feature set unique and less vulnerable to multicollinearity.
- Correlation Coefficient plays an important role in Factor analysis method.
- A classification technique, factor analysis can be used a dimensionality reduction.
- Here groups are created by combining highly correlated feature where the groups are not correlated with each other. 2 main types - i) EFA ii) CFA

3) LDA (Linear Discrimination Analysis) →

- LDA is a dimensionality reduction technique. Goal of LDA is to project the feature in higher dimensional space onto a lower dimensional space in order to avoid the curse of dimensionality
- So first LDA need to calculate the seperability between classes which is distance between the mean of different classes. This is called between class variance.
- Secondly calculate the distance between mean and sample of each class. It is also called within class variance.
- Finally construct the lower dimensional space which maximize the between class variance & minimize the within class variance.

4) T-Sne

- T- distributed Stochastic Neighbour Embedding is a non-linear dimensionality reduction algorithim used for exploring high dimensional data.
- t-SNE finds pattern in the data by identifying observed clusters based on similarity of data points with multiple feature. It maps the multi-dimensional data to lower dimensional space, input feature are no longer identifiable.
- T-SNE can be used in process of classification and clustering by using its output as the input feature for other classification algorithim.